A taxonomy of IRT models for ordening persons and items using simple sum scores

Sijtsma, K.; Hemker, B.T.

Published in:
Journal of Educational and Behavioral Statistics

Publication date:
2000

Document Version
Publisher's PDF, also known as Version of record

Journal of Educational and Behavioral Statistics
Winter 2000, Vol. 25, No. 4, pp. 391–415

A Taxonomy of IRT Models for Ordering Persons and Items Using Simple Sum Scores

Klaas Sijtsma
Tilburg University

Bas T. Hemker
CITO National Institute for Educational Measurement

Keywords: *dichotomous IRT models, invariant item ordering, item ordering, item response theory, person ordering, polytomous IRT models, stochastic ordering*

The stochastic ordering of the latent trait by means of the unweighted total score is considered for 10 dichotomous IRT models and 10 polytomous IRT models. The conclusion is that the stochastic ordering property holds for all dichotomous IRT models and for two polytomous IRT models. Also, the invariant item ordering property is considered for the same 20 IRT models. It is concluded that invariant item ordering holds for three dichotomous IRT models and three polytomous IRT models. The person and item ordering results are summarized in a taxonomy of IRT models. Some consequences for practical test construction are briefly discussed.

Item response theory (IRT) makes a sharp distinction between the observable scores of a respondent on a set of items and the scale on which the unobservable psychological construct is measured. The construct can be a personality trait, a cognitive ability, an educational achievement, an attitude, or an opinion, in short, a latent trait. Typical of IRT measurement is that interest almost always lies with the respondent's position on the latent trait scale or, simply, the latent trait, to be denoted θ. The observable scores on a set of well-chosen items are used to estimate θ. When abilities or achievements are measured, the item scores may reflect whether answers are correct or incorrect, or may reflect degrees of correctness, depending on the types of errors made. When personality traits or attitudes are measured, item scores may reflect the degree of endorsement with a particular statement. Response categories may be labeled options such as "Never," "Rarely," "Occasionally," "Often," and "Always," or another labeling, which depends on the wording of the item. Usually, higher item scores are assumed to reflect a higher θ.

More specifically, sum scores or other functions of the item scores are used to estimate a person's θ value. For example, in the Rasch (1960) model or 1-parameter logistic model (1PLM), the unweighted sum of the item scores, denoted X_+, is a sufficient statistic for estimating θ. In the 2-parameter logistic model (2PLM; Birnbaum, 1968), a weighted sum of item scores, with each item weighted by its discrimination parameter, is the sufficient statistic for estimating

θ. Since the discrimination parameters usually are unknown, in practice in the 2PLM and the 3-parameter logistic model (3PLM; Birnbaum, 1968) a respondent's pattern of 1s for correct responses and 0s for incorrect responses is used to estimate θ. Estimates of θ may be used to order the respondents on θ and, for example, to select the 20 respondents with the highest $\hat{\theta}$s for an expensive follow-up course or all respondents with $\hat{\theta}$s below a preset level θ_{cutoff} for remedial teaching.

Although in an IRT context θ is used for measuring persons and X_+ seems to be useful primarily for estimating θ in, e.g., the 1PLM, classical sum scores such as X_+ have not lost their practical value for measuring persons. We will argue that since the interpretation of θ is rather complex and remote from everyday experience, the θ scale may not be convenient for communicating test performance results to measurement practitioners (such as test constructors and psychologists who administer tests) and their clients (such as organizations and government institutions and the individuals tested at their request) and to pupils and their teachers and parents. Although the interpretation of θ is evident for psychometricians who understand the concepts of probability, odds, and logit, test users and their clients are not familiar with these concepts.

As an example of the complexity of θ we will consider the difference between two θs under the relatively simple 1PLM. Let X_i be the random variable denoting the score on item i, with realizations $x = 0, 1$; let $P_i(\theta) \equiv P(X_i = 1|\theta)$; and let δ_i be a latent location parameter; then the item response function (IRF) of the 1PLM can be defined as

$$P_i(\theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)}. \tag{1}$$

It may be noted that the odds of a fixed θ_v producing a 1 score on item i, denoted O_{vi}, is $O_{vi} = \exp(\theta_v - \delta_i)$. It follows readily that under the 1PLM (Equation 1) the difference between two latent trait values, say θ_v and θ_w, can be expressed as a difference in logits,

$$\theta_v - \theta_w = \ln(O_{vi}) - \ln(O_{wi}) = logit[P_i(\theta_v)] - logit[P_i(\theta_w)]. \tag{2}$$

Clearly, Equation 2 is highly technical, and for most non-psychometricians the difference between θs probably will not have a close reference to the real world.

One could argue, of course, that psychometricians should educate or explain the meaning of equations such as (2) to measurement practitioners and laymen. We think this approach may be successful for those few members of these groups who already have a basic knowledge of psychometrics. For the other members a more fruitful approach to communicating test results may be the use of simpler scales, and then X_+ is a likely candidate. This is not to say that X_+ or another summary score based on observable item scores have a self-evident meaning, but those scores are more familiar due to their direct relation to "doing things right or wrong" (abilities, achievements) or "earning more or fewer points" (personality traits, attitudes) than θ scores which lack such a direct relation.

If for practical reasons one would use, say, X_+ for person measurement rather than θ, some psychometricians might argue that within an IRT framework this is bad practice because in many IRT models $\hat{\theta}$ uses more information from the pattern of item scores than X_+ and that, moreover, the use of the test information function (Lord, 1980, pp. 65–80) provides a better evaluation of measurement precision by means of $\hat{\theta}$ than the classical reliability coefficient does for X_+. However, nothing prevents the simultaneous use of IRT for test construction and the information function for measurement evaluation of $\hat{\theta}$ on the one hand, and the communication of test performance to measurement practitioners and laymen by means of scores such as X_+ on the other hand. Thus, there still may be a role for X_+ for measuring persons, in particular, for communicating measurement results. In fact, this is what often happens in practical use of tests.

In what follows, we will consider $X_+ = \Sigma X_i$, with $X_i = 0, 1, \ldots, m$; thus, X_+ is defined both for dichotomous and polytomous items. For dichotomous items X_+ is a natural candidate, but for polytomous items summary scores based on the weighting of the item scores might replace X_+. Because in practice researchers often use Likert scoring for their rating scales, the unweighted X_+ seemed to be a reasonable choice.

The first goal of this paper is to show that for a broad class of IRT models for dichotomous items the use of X_+ for measuring persons is a sensible practice, but that one runs into trouble as soon as X_+ is based on ordered polytomous item responses. We will discuss, in particular, that under several of the well known polytomous IRT models a higher X_+ value does *not* always imply a higher θ value; that is, X_+ does not *stochastically order* θ (Hemker, Sijtsma, Molenaar, & Junker, 1997). This means that with several polytomous IRT models an ordering of the respondents on X_+ can give a misleading impression about their ordering on θ. Thus, the well known and long-appreciated sum score X_+ can be misleading for ordering respondents in a polytomous IRT context.

In addition to the ordering of persons, test constructors are often interested in the ordering of their items. In several applications of a test, it may be desirable to have the same ordering of the items in different subgroups, for example, according to gender, ethnic or social background, or previous educational level. This may be relevant in differential item functioning (DIF; e.g., Holland & Wainer, 1993) or test fairness research. Sometimes it may be desirable to have the same item ordering for each θ value as, for example, in person-fit research (e.g., Meijer, Molenaar, & Sijtsma, 1994).

As in the case of person measurement, it would be convenient for measurement practitioners to use a simple and familiar statistic for ordering the items. We will use the mean item score conditional on θ, that is, $E(X_i|\theta)$, where the expectation runs across all respondents with the same θ. Again, one might argue that, since IRT provides estimates of latent location parameters δ as in Equation 1, these estimates should be used for describing item difficulty. Moreover, the maximum likelihood estimates of these location parameters often provide the maximum Fisher information.

Three remarks about δ are worth considering. First, as with θ, the use of δ would introduce problems in communicating test performance results to practitioners and their clients. Second, in dichotomous IRT models, such as the 1-, 2-, and 3PLM, the maximum likelihood estimate of δ provides maximum Fisher information, but this is *not* generally true in polytomous IRT models, e.g., the partial credit model (Masters, 1982). For example, for partial credit model items with three answer categories, described by two item location parameters, δ_1 and δ_2, Muraki (1993, p. 356) provided graphs of unimodal and bimodal *item category* information functions. Akkermans and Muraki (1997) showed that for three-category items the *item* information function is unimodal if $\delta_2 - \delta_1 < 4/ln2$ and bimodal otherwise. Thus, in the partial credit model there is no one-to-one correspondence between location parameters and modes of information functions. Third, contrary to what may be tempting to believe, in most IRT models δ *cannot* be interpreted as the difficulty of an item (dichotomous items) or an item answer category (polytomous items; Molenaar, 1983; Verhelst & Verstralen, 1991). For example, consider the 2PLM, in which α_i denotes the slope parameter,

$$P_i(\theta) = \frac{\exp[\alpha_i(\theta - \delta_i)]}{1 + \exp[\alpha_i(\theta - \delta_i)]}; \alpha_i > 0. \tag{3}$$

The IRFs of two items i and j with $\alpha_i \neq \alpha_j$, intersect at

$$\theta_{ij} = \frac{\alpha_i\delta_i - \alpha_j\delta_j}{\alpha_i - \alpha_j}. \tag{4}$$

Arbitrarily, assume that $\alpha_i < \alpha_j$; then for each $\theta < \theta_{ij}$ we have that $P_i(\theta) > P_j(\theta)$, and for each $\theta > \theta_{ij}$ we have that $P_i(\theta) < P_j(\theta)$. Thus, for an examinee with a $\theta < \theta_{ij}$ the subjective (i.e., given θ) response probabilities show that item i is easier than item j, and for another examinee with a $\theta > \theta_{ij}$ the item ordering is reversed. The difficulty ordering is reflected by the conditional response probabilities and for items with given δs this ordering depends on θ (in fact, δ gives the location of the inflection point of a logistic curve).

For polytomous IRT models, δ is even more remote from being a difficulty parameter. For example, in the partial credit model (Masters, 1982) each item with $m + 1$ ordered answer categories has m transition parameters (Masters, 1982) or threshold parameters (Andrich, 1995), denoted δ_{ix} ($x = 1, \ldots, m$). The category characteristic curve (CCC), $P(X_i = x|\theta)$, is defined as

$$P(X_i = x|\theta) = \frac{\exp[\sum_{s=1}^{x}(\theta - \delta_{is})]}{\sum_{q=0}^{m}\exp[\sum_{s=1}^{q}(\theta - \delta_{is})]}. \tag{5}$$

The distances between the δ_{ix}s of an item are not fixed across items and the ordering of the δ_{ix}s may vary for different items (Masters, 1982). The δ_{ix} parameter gives the location of the intersection point of the CCCs of the categories x-1 and x. The δ_{ix} parameters do not provide information on the ordering of conditional response probabilities, and combinations of the m location parameters of each of the items do not provide information on the ordering of the items.

We conclude that in dichotomous IRT models the location parameter δ is not an unequivocal difficulty parameter when IRFs cross and that in the partial credit model (polytomous items) location parameters indicate intersection points of CCCs of adjacent categories and, moreover, that for this and other polytomous IRT models no *item* difficulty parameter exists. Thus, a simple statistic that expresses item difficulty, which also is useful for communicating test results to researchers, is badly needed. $E(X_i|\theta)$ is such a statistic. It provides information about difficulty at the item level and it does this for varying θs.

The second goal of this paper is to discuss the fact that only a limited number of IRT models imply an item ordering according to conditional item means $E(X_i|\theta)$ that is the same with the exception of possible ties, for all θs. Such an ordering is an *invariant item ordering* (IIO; Sijtsma & Junker, 1996; Sijtsma & Hemker, 1998). For IRT models for dichotomous items it is fairly simple to see which models do and which models do not imply an IIO, but for polytomous IRT models neither of these options is obvious. In fact, most polytomous IRT models do not imply an IIO. Thus, a researcher interested in such an ordering cannot rely on most IRT models to produce it.

This paper is concerned with the relation between IRT and classical statistics for measurement of persons and items. Considered this way, the paper fits into a tradition of research that addresses relations between modern (IRT) and classical test theory. For example, Mokken (1971, pp. 142–147) proposed a classical reliability coefficient based on a nonparametric IRT model; Lord (1980, pp. 33–43) discussed the relations between item parameters from the 1PLM, 2PLM, and 3PLM framework and classical test theory; and Mellenbergh (1996) discussed the relations between classical reliability and the information function used for evaluating measurement precision in IRT.

Based on theoretical work of Grayson (1988), Hemker et al., (1996, 1997), Huynh (1994), Sijtsma and Junker (1996), and Sijtsma and Hemker (1998), and on some new results to be presented here, we discuss a taxonomy of 10 IRT models for dichotomous items and 10 IRT models for polytomous items, which shows which models imply (1) the correct ordering of persons on θ by means of X_+ (stochastic ordering of θ using X_+; abbreviated SOL) and (2) an ordering of the items by conditional item means $E(X_i|\theta)$ that is the same in all possible subgroups (IIO). Moreover, we discuss practical consequences for person and item measurement when models lacking one or both ordering properties are used to analyze one's data.

Using X_+ for Stochastic Ordering of Persons on the Latent Trait

The IRF typically is assumed to be monotonely nondecreasing in θ, indicating that a higher θ level means a higher probability of giving the correct answer (dichotomous items) or a higher probability of obtaining at least x points on a rating scale (polytomous items). The IRF can be a logistic function (equations 1 and 3) or a normal-ogive function (examples are given later on), but other choices also are possible, such as a logistic function to the power of ξ ($\xi \neq 1$; the result is not a logistic function) as in the acceleration model (Samejima, 1995). Similarly, Agresti (1990, ch. 9) discusses several ordinal response models using logit link functions and cumulative link models using probit link functions and complementary log-log link functions. Given the monotonicity requirement of IRT, the choice of an IRF (or, similarly, a link function) is based on statistical or data-based criteria. For example, the choice of the 1-parameter logistic function is based on the (minimal) sufficiency of X_+ for the estimation of θ and the item total score for the estimation of δ (Molenaar, 1995). Van Engelenburg (1997) argues that the choice of a particular polytomous IRT model for analyzing one's data should be governed by the type of item used, and Akkermans (1998) argues that the scoring rule of the items should determine the polytomous IRT model to be used.

A consequence of the choice of a particular IRF or a particular link function is that the measurement properties of the model are determined at least partly by this choice. An example is the property of specifically objective measurement, typical of the 1PLM (Equation 1) and the 2PLM (Equation 3) (Irtel, 1995). Another example is the difference scale level of θ in the 1PLM (Equation 1) and the interval scale level of θ in the 2PLM (Equation 3). Alternatively, we will not choose a priori a particular parametric IRF implying certain measurement properties; rather, we will start from the measurement property of stochastic ordering, which can be seen as a general practical requirement for measurement models comparable with the monotonicity of the IRF. After a discussion of the stochastic ordering property, the next step is to investigate which IRT models imply this property. First, we explain the stochastic ordering property.

Consider two total scores, $X_+ = s_1$, s_2, and assume that $s_1 < s_2$. We require throughout that a group of individuals with $X_+ = s_2$ should have a higher mean θ than a group with $X_+ = s_1$. This requirement follows from an assumption about quantities closely related to the cumulative distributions of θ in these total score groups, which can be formalized as follows. Let each of the k items in the test be scored in the same way, for example 0–1, or 0–1–2–3–4. The assumption is that the probability that θ is at least some arbitrary constant t is nondecreasing in X_+; thus, for two total scores, s_1 and s_2, with $s_1 < s_2$,

$$P(\theta \geq t \,|\, X_+ = s_1) \leq P(\theta \geq t \,|\, X_+ = s_2). \tag{6}$$

Equation 6 says that the latent trait θ is *stochastically ordered* by X_+ (SOL; Hemker et al., 1996, 1997; also see Lehmann, 1959, 1994, p. 84, and

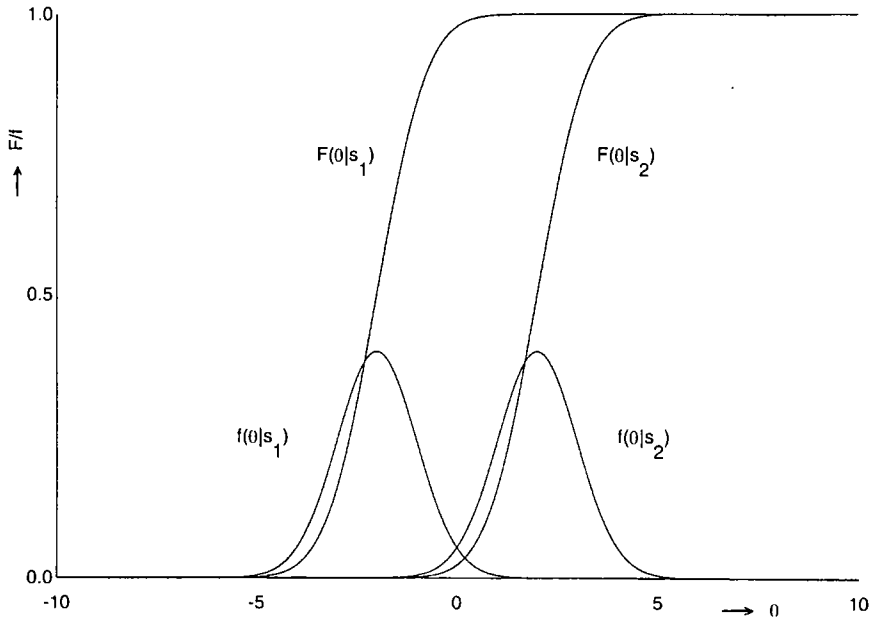

FIGURE 1. *Two Cumulative Normal Distributions of* θ *for* $X_+ = s_1$ *[Left;* $F(\theta|s_1)$*] and* $X_+ = s_2$ *[Right;* $F(\theta|s_2)$*]; and Corresponding Probability Density Functions,* $f(\theta|s_1)$ *and* $f(\theta|s_2)$

Bartholomew, 1996, pp. 169–171, citing Knott & Albanese, 1993, who discusses the identical ordering of X_+ and $E(\theta|X_+)$).

Another way to look at Equation 6 is in terms of conditional cumulative distribution functions of θ for s_1 and s_2. Subtracting both sides in Equation 6 from 1 yields

$$P(\theta \leq t|X_+ = s_1) \geq P(\theta \leq t|X_+ = s_2). \tag{7}$$

Thus, the cumulative distribution function of θ is uniformly larger for s_1 than for s_2. This is displayed in Figure 1 for two cumulative normal distributions, denoted $F(\theta|s_1)$ and $F(\theta|s_2)$, together with the corresponding probability density functions, denoted $f(\theta|s_1)$ and $f(\theta|s_2)$. Obviously, for the group with the smaller $X_+ = s_1$ the mean of θ is smaller than for the group with $X_+ = s_2$. Equation 6 does *not* hold for each IRT model. Such models thus may provide little confidence in the ordering of respondents on θ by means of X_+. In this paper we present SOL results for IRT models that were obtained without assuming a particular prior distribution of θ (based on theoretical results by Grayson, 1988, and Hemker et al., 1996, 1997; see also Bartholomew, 1996, p. 170).

Invariant Item Ordering

In general, an ordering of items that is the same (except for possible ties) for all θs facilitates the interpretation of test results. Such an ordering is an IIO (see Sijtsma & Junker, 1996, for dichotomous items and Sijtsma & Hemker, 1998, for polytomous items). An IIO holds if the k items can be ordered and numbered such that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \ldots \leq E(X_k|\theta); \textit{for all } \theta; X_i = 0, \ldots, m. \tag{8}$$

Equation 8 says that for each value of θ the ordering of the item means is the same, except for possible ties. From Equation 8 follows that the same item ordering also holds in each subgroup from the population of interest. If the item scores are equal to 0 or 1, we know that $E(X_i|\theta) = P(X_i = 1|\theta)$; that is, the conditional mean item score equals the conditional probability of giving the correct answer to the item. This is the IRF for dichotomous items.

Assumptions and Distinctions

The dichotomous IRT models and polytomous IRT models discussed here have three common assumptions. The first assumption is unidimensionality (UD), which means that all items in the test measure the same trait. Mathematically, UD means that only one person parameter θ accounts for the data structure. Thus, θ is a scalar.

Local independence (LI), which is the second assumption, means that the response of an individual to an item from the test is not influenced by his or her responses to the other items from that same test or by other traits than θ. Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ be the vector that contains the item score random variables, and let $\mathbf{x}$ denote a realization of $\mathbf{X}$ which contains k numerical item scores. UD and LI mean that

$$P(\boldsymbol{X} = \boldsymbol{x}|\theta) = \prod_{i=1}^{k} P(X_i = x_i|\theta). \tag{9}$$

Integrating Equation 9 across the distribution of θ yields the manifest distribution $P(\mathbf{X} = \boldsymbol{x})$. Let the cumulative distribution function of θ be denoted $F(\theta) = P(\theta \leq t)$. For the moment, we only consider tests consisting of dichotomous items; thus, we write $P_i(\theta) \equiv P(X_i = 1|\theta)$, for short. Integrating across θ yields

$$P(\boldsymbol{X} = \boldsymbol{x}) = \int_{\theta} \prod_{i=1}^{k} P_i(\theta)^{x_i} \cdot [1 - P_i(\theta)]^{1-x_i} dF(\theta). \tag{10}$$

For polytomous items, an equation for $P(\mathbf{X} = \boldsymbol{x})$ can be obtained by integrating the righthand side of Equation 9 across θ. It has been noted (Holland & Rosenbaum, 1986; Suppes & Zanotti, 1981) that Equation 10 does not restrict the data unless there are additional assumptions on the $P_i(\theta)$s, or on $F(\theta)$, or on both. In IRT, the assumptions usually pertain to the IRFs.

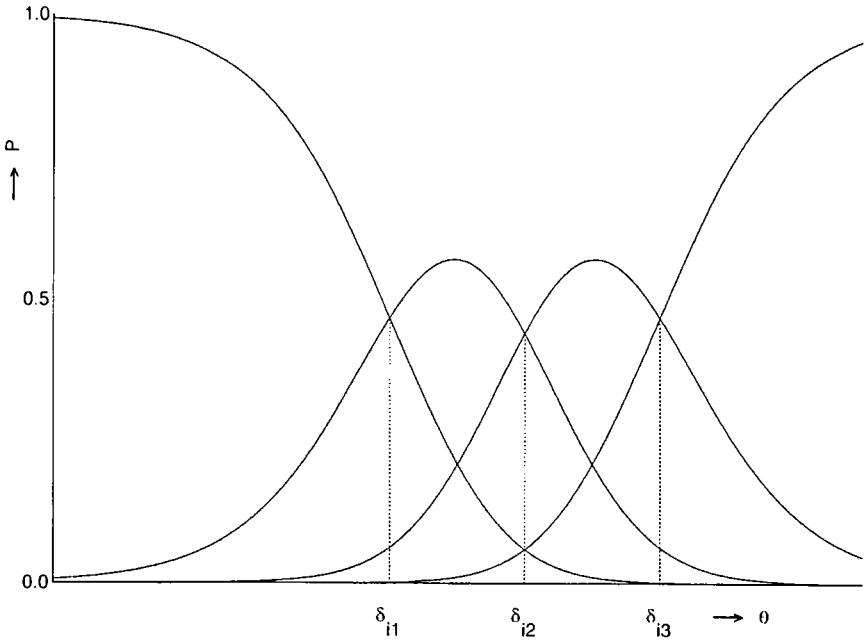

FIGURE 2. *Four CCCs of a Four-Category Item Under the Partial Credit Model*

For *dichotomous* items the IRF is assumed to be monotonely nondecreasing in θ or strictly increasing in θ. We use the abbreviation M for monotonicity to capture both conditions. M is the third assumption.

Because IRT models for *polytomous* items based on UD and LI have $m + 1$ ordered answer categories (scored $x = 0, 1, \ldots, m$), for each item m functions $P(X_i \geq x|\theta)$ $(x = 1, \ldots, m)$ are used to describe the relation between X_i and θ. These functions are the item step response functions (ISRFs; e.g., Hemker et al., 1997). Some polytomous IRT models, such as the partial credit model (Masters, 1982) rather define the CCC, $P(X_i = x|\theta)$; see Equation 5, but this probability easily can be coverted to $P(X_i \geq x|\theta)$, and vice versa (e.g., Sijtsma & Hemker, 1998), because

$$P(X_i = x|\theta) = P(X_i \geq x|\theta) - P(X_i \geq x + 1|\theta), \tag{11}$$

and

$$P(X_i \geq x|\theta) = \sum_{y=x}^{m} P(X_i = y|\theta). \tag{12}$$

For a four-category item, Figure 2 shows four typical CCCs; and for $x = 1, 2, 3$, Figure 3 shows three typical ISRFs.

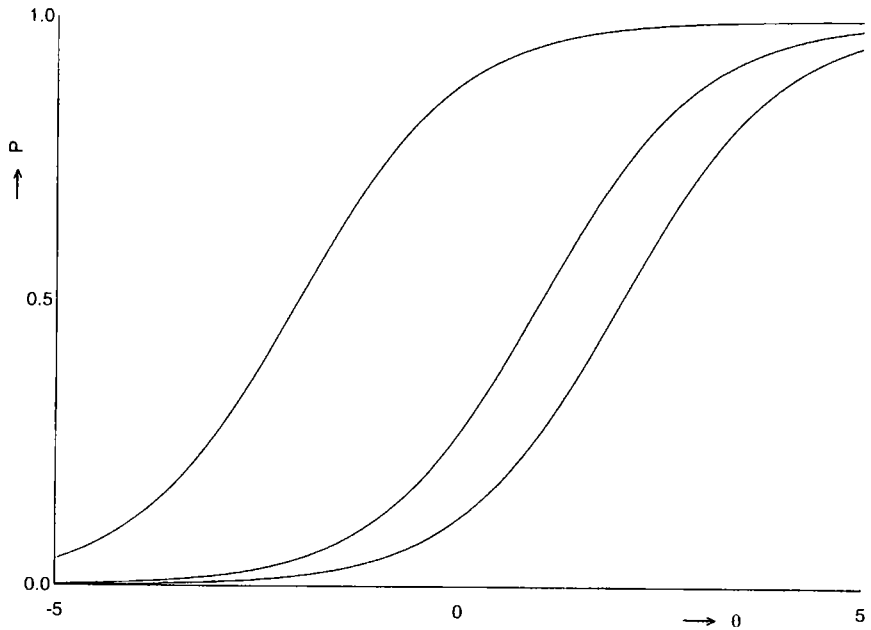

FIGURE 3: *Three ISRFs of a Four-Category Item Under the Graded Response Model*

Based on the ISRFs, for polytomous IRT models *item* response functions also can be defined (Chang & Mazzeo, 1994). For item i ($X_i = 0, \ldots, m$), the IRF is defined (Sijtsma & Hemker, 1998) as the sum of the m ISRFs,

$$E(X_i|\theta) = \sum_x xP(X_i = x|\theta) = \sum_x P(X_i \geq x|\theta). \quad (13)$$

Note that like the IRF for dichotomous items, the IRF for polytomous items is the conditional expected item score, but unlike the IRF for dichotomous items, the IRF for polytomous items is not a probability. Specifically, its range is $0 \leq E(X_i|\theta) \leq m$. Note that $E(X_i|\theta)$ is used for defining an IIO; see Equation 8.

IRT Models for Dichotomous Item Scores

The assumptions of UD, LI, and M together define a class of several popular and much used IRT models. We distinguish ten models in total; eight parametric IRT models, and two nonparametric IRT models.

Parametric IRT Models

The first model is the 1PLM (Rasch, 1960) with varying location or difficulty parameters δ for the items and IRF defined in Equation 1. The IRFs of the 1PLM are parallel functions. The second model is the 2PLM (Birnbaum, 1968)

with varying location, and varying slope or discrimination parameters α and IRF defined in Equation 3. The IRFs of the 2PLM are allowed to intersect (Equation 4). The third model is the 3PLM (Birnbaum, 1968) with varying location, varying slope, and varying lower asymptote or guessing parameters γ; the IRF is defined as

$$P_i(\theta) = \gamma_i + \frac{(1 - \gamma_i)\exp[\alpha_i(\theta - \delta_i)]}{1 + \exp[\alpha_i(\theta - \delta_i)]}; 0 < \gamma_i < 1. \tag{14}$$

The next three models are the well known 1-, 2-, and 3-parameter normal ogive models (e.g., Lord, 1952, 1980) with the same types of item parameters as their logistic counterparts. These three models can be seen as the historical predecessors of the logistic models, which have more convenient mathematical properties and thus have replaced the normal ogive models in practice. For completeness we provide the model equations of the normal ogive models. The parameters δ, α, and γ have been replaced by b, a, and c, respectively, which have the same interpretation as δ, α, and γ. The 1-, 2-, and 3-parameter normal ogive models are given by

$$P_i(\theta) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{(2\pi)}} e^{-z_i^2/2} dz; z_i = \theta - b_i; \tag{15}$$

$$P_i(\theta) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{(2\pi)}} e^{-z_i^2/2} dz; z_i = a_i(\theta - b_i); \tag{16}$$

and

$$P_i(\theta) = c_i + (1 - c_i) \int_{-\infty}^{z_i} \frac{1}{\sqrt{(2\pi)}} e^{-z_i^2/2} dz; z_i = a_i(\theta - b_i), \tag{17}$$

respectively. Although mathematically different, the normal ogive IRFs and the logistic IRFs have almost the same shape. The maximum resemblance is obtained if the exponent in the numerator and the denominator of the Equations 1, 3, and 14 is multiplied by a constant D = 1.7 (Hambleton & Swaminathan, 1985, p. 37).

The logistic models and the normal ogive models are well-known. Two rather unknown parametric models are the 4-parameter logistic model (4PLM; e.g., Hambleton & Swaminathan, 1985, p. 48) and the One-Parameter Logistic Model with imputed slopes (OPLM; Verhelst & Glas, 1995).

In addition to a location parameter δ, a slope parameter α, and a lower asymptote γ, the 4PLM has an upper asymptote parameter ζ. The IRF is a natural generalization of the 3PLM, as it is defined as

$$P_i(\theta) = \gamma_i + \frac{(\zeta_i - \gamma_i)\exp[\alpha_i(\theta - \delta_i)]}{1 + \exp[\alpha_i(\theta - \delta_i)]}; \gamma_i < \zeta_i < 1. \tag{18}$$

The 4PLM is mentioned here because it allows items to be difficult in the sense that even the most able examinees have a non-trivial probability of failing. This makes the model conceptually interesting. Unfortunately, the 4PLM has little practical relevance because it has too many parameters to be estimated.

The OPLM is mentioned because it combines the statistical virtues of the 1PLM with the greater flexibility of the 2PLM. This is accomplished by the imputation of integer values for the slope parameters rather than the statistical estimation of these parameters. Let the slope index be denoted A_i, then the IRF is

$$P_i(\theta) = \frac{\exp[A_i(\theta - \delta_i)]}{1 + \exp[A_i(\theta - \delta_i)]}; A_i \in N^+. \tag{19}$$

As a result, only the location or difficulty parameters are estimated and the imputed slopes may be adapted in consecutive iterations until a satisfactory fit of the model to the data is obtained.

Each of the models in Equations 1, 3, and 14 through 19 parametrically defines the IRF by means of either the logistic or the normal ogive function. Hence these are *parametric* IRT models.

Nonparametric IRT Models

Nonparametric IRT models place order restrictions on the IRFs but refrain from a parametric definition. Two models that have frequently been used for test and questionnaire construction are the model of monotone homogeneity (MHM; Mokken & Lewis, 1982; Sijtsma, 1998) and the model of double monotonicity (DMM; Mokken & Lewis, 1982; Sijtsma & Junker, 1996; Sijtsma, 1998). The MHM assumes that the IRFs are monotonely nondecreasing. This means that for any pair of θs with $\theta_v < \theta_w$,

$$P_i(\theta_v) \leq P_i(\theta_w). \tag{20}$$

Note that the MHM is defined completely by UD, LI, and M. The MHM can be seen as a nonparametric version of the 3PLM (Equation 14) or the 4PLM (Equation 18), or the 3-parameter normal ogive model (Equation 17).

The DMM assumes that the IRFs are monotonely nondecreasing and, in addition, that they do not intersect. This means that for two items i and j, if we know that for one θ, $P_i(\theta) < P_j(\theta)$, then

$$P_i(\theta) \leq P_j(\theta), \textit{for all } \theta. \tag{21}$$

The DMM can be seen as a nonparametric version of the 1PLM (Equation 1) and the 1-parameter normal ogive model (Equation 15) (Meijer, Sijtsma, & Smid, 1990; Sijtsma, 1998).

Variation on M in Dichotomous IRT Models

To summarize, all models mentioned have UN, LI, and M in common, and differ in the additional restrictions placed on M. These restrictions can be

TABLE 1
Characteristics of IRFs of 10 IRT Models for Dichotomous Items

	Lowest Prob.	Highest Prob.	Slope	#Inflection Points	Intersection
1-PLM	0	1	const	1	No
2-PLM	0	1	var	1	Yes
3-PLM	>0	1	var	1	Yes
4-PLM	>0	<1	var	1	Yes
OPLM	0	1	var	1	Yes
1-PNOM	0	1	const	1	No
2-PNOM	0	1	var	1	Yes
3-PNOM	>0	1	var	1	Yes
MHM-di	≥0	≤1	var	≥1	Yes
DMM-di	≥0	≤1	var	≥1	No

List of Abbreviations:
1-PLM: 1-Parameter Logistic Model (Rasch model)
2-PLM: 2-Parameter Logistic Model (Birnbaum model)
3-PLM: 3-Parameter Logistic Model
4-PLM: 4-Parameter Logistic Model
OPLM: One-Parameter Logistic Model with Imputed Slopes
1-PNOM: 1-Parameter Normal Ogive Model
2-PNOM: 2-Parameter Normal Ogive Model
3-PNOM: 3-Parameter Normal Ogive Model
MHM-di: Model of Monotone Homogeneity for Dichotomous Data
DMM-di: Model of Double Monotonicity for Dichotomous Data

characterized by parametric and nonparametric definitions of the IRF. Alternatively, the variations on M are summarized in Table 1 and can be described as pertaining to the following:

(1) The lowest value of the IRF. This is not 0 in the 3PLM, the 4PLM, and the 3-parameter normal ogive model, and not necessarily 0 in the MHM and the DMM. Thus each of these models allows nonzero probabilities for low θs due, e.g., to guessing. The lower asymptote equals 0 in the other models.

(2) The highest value of the IRF. This is not 1 in the 4PLM, and not necessarily 1 in the MHM and the DMM. These models allow the possibility of failure even for very high θs. The upper asymptote equals 1 in the other parametric models.

(3) The slope of the IRF. The slopes are equal for all items in the 1PLM and the 1-parameter normal-ogive model, but they may vary in all other models.

(4) The inflection point of the IRF. For the four logistic models, the three normal-ogive models, and the OPLM, this point is located exactly at the item location on the θ scale. For the MHM and the DMM there is not a fixed location, and one IRF even can have several inflection points.

(5) Intersection of the IRFs. This is not possible in the 1PLM, the 1-parameter normal ogive model, and the DMM. It is allowed in all other models.

IRT Models for Polytomous Item Scores

Like dichotomous models, polytomous models can be parametric or nonparametric, depending on whether the CCCs or ISRFs are parametrically defined or whether these functions are only subject to order restrictions. Within the class of parametric models, we follow Thissen and Steinberg (1986) in distinguishing divide-by-total models and difference models.

Parametric Polytomous IRT Models: Divide-By-Total Models

The first model to be considered is the well known partial credit model (Masters, 1982). The partial credit model parametrically defines the CCC; see Equation 5. There are no restrictions on the distances between the locations of the CCCs of one item. The second model is the generalized partial credit model (Muraki, 1992). Compared with the partial credit model, this model has a slope parameter α_i, which is fixed for all CCCs of item i, as is shown by

$$P(X_i = x|\theta) = \frac{\exp[\sum_{s=1}^{x} \alpha_i(\theta - \delta_{is})]}{\sum_{q=0}^{m} \exp[\sum_{s=1}^{q} \alpha_i(\theta - \delta_{is})]}. \tag{22}$$

The third model is the rating scale model (Andrich, 1978); this is a special case of the partial credit model in that it is assumed that $\delta_{ix} = \delta_i + \tau_x$; δ_i is a location parameter, and the thresholds are characterized by m parameters τ_x ($x = 1, \ldots, m$). The item parameter δ_i is defined as the mean of the δ_{ix}s across x. The CCC is defined as

$$P(X_i = x|\theta) = \frac{\exp[\sum_{s=1}^{x} (\theta - \delta_i - \tau_s)]}{\sum_{q=0}^{m} \exp[\sum_{s=1}^{q} (\theta - \delta_i - \tau_s)]}. \tag{23}$$

Patterns of corresponding τs of different items i and j can be obtained through translations equal to $\delta_i - \delta_j$. Note that the partial credit model and the generalized partial credit model relate like the 1PLM and the 2PLM. For dichotomous items, the rating scale model reduces to the 1PLM.

The OPLM for polytomous items with imputed slopes (Verhelst & Glas, 1995) is a hybrid between the partial credit model and the generalized partial

credit model because the slope parameters are imputed and only location parameters are statistically estimated. The CCC is defined as

$$P(X_i = x|\theta) = \frac{\exp[\sum_{s=1}^{x} A_i(\theta - \delta_{is})]}{\sum_{q=0}^{m} \exp[\sum_{s=1}^{q} A_i(\theta - \delta_{is})]}; A_i \in N^+. \quad (24)$$

The models defined by the Equations 5, and 22 through 24 are divide-by-total models (Thissen & Steinberg, 1986; Hemker et al., 1997).

Parametric Polytomous IRT Models: Difference Models

The next two models are difference models (Thissen & Steinberg, 1986; Hemker et al., 1997). The first is the graded response model (Samejima, 1969), which parametrically defines the ISRF, $P(X_i \geq x|\theta)$. Within the same item, the ISRFs have a fixed order (which is always true; see Equation 12), parameterized by m threshold parameters with $\lambda_{i1} \leq \lambda_{i2} \leq \ldots \leq \lambda_{im}$. The distances between adjacent ISRFs of the same item are free to vary. The ISRF is defined as

$$P(X_i \geq x|\theta) = \frac{\exp[\alpha_i(\theta - \lambda_{ix})]}{1 + \exp[\alpha_i(\theta - \lambda_{ix})]}; \alpha_i > 0. \quad (25)$$

The relative position of the ISRFs of different items is not restricted. The rating scale version of the graded response model (Muraki, 1990) is a special case of the graded response model in that it restricts the location parameter. Let λ_i denote the location parameter of item i, and β_x a parameter of the *x-th* ISRF. By assuming that $\lambda_{ix} = \lambda_i + \beta_x$, the ISRF of the rating scale version of the graded response model is

$$P(X_i \geq x|\theta) = \frac{\exp[D\alpha_i(\theta - \lambda_i - \beta_x)]}{1 + \exp[D\alpha_i(\theta - \lambda_i - \beta_x)]}, \quad (26)$$

where D is a scaling constant that puts the θ-scale in the same metric as the normal ogive model.

Nonparametric Polytomous IRT Models

In the class of nonparametric IRT models we consider four models, which are all difference models (Thissen & Steinberg, 1986; Hemker et al., 1997). The MHM (Molenaar, 1982, 1997) is a nonparametric version of Samejima's graded response model (Equation 25); Hemker et al. (1997) therefore call the MHM the nonparametric graded response model. The MHM assumes that the ISRF is a nondecreasing function of θ; thus, for any pair $\theta_v < \theta_w$,

$$P(X_i \geq x|\theta_v) \leq P(X_i \geq x|\theta_w), \textit{for all } i; \textit{and for all } x. \quad (27)$$

Hemker et al. (1997) showed that the parametric divide-by-total models (Equations 5; 22 through 24) and the parametric difference models (Equations 25 and 26) are all special cases of the MHM for polytomous items (Equation 27). The classes of parametric divide-by-total models and parametric difference models are mutually exclusive, however (Thissen & Steinberg, 1986). A consequence of the MHM being the most general model is that all models discussed so far (and also all nonparametric models to be discussed next) have nondecreasing ISRFs; thus, all polytomous IRT models have ISRFs that are M (also see Sijtsma & Hemker, 1998, Lemma).

Within the class of nonparametric polytomous IRT models, we mention three more models, which are all special cases of the MHM. The three models are all characterized by nonintersection of all ISRFs or of subsets of ISRFs. The first model is the DMM (Molenaar, 1982, 1997), which assumes that, in addition to M (Equation 27), the ISRFs of different items do not intersect. Thus, for two items i and j, if we know that for one θ_v, $P(X_i \geq s|\theta_v) < P(X_j \geq r|\theta_v)$, then

$$P(X_i \geq s|\theta) \leq P(X_j \geq r|\theta), \textit{for all } \theta; \textit{ and for all } s,r. \tag{28}$$

Equation 28 can be extended to k items. The DMM thus allows any ordering of ISRFs across items, given the structural restriction that the ordering within items is fixed (see Equation 12). The second model is the strong DMM (Sijtsma & Hemker, 1998); in addition to M (Equation 27) and nonintersection of the ISRFs (Equation 28), this model assumes that for given item score x,

$$P(X_i \geq x|\theta) \leq P(X_j \geq x|\theta), \textit{for all } \theta; \textit{ and for all } x. \tag{29}$$

Sijtsma and Hemker (1998) called the DMM the weak DMM to distinguish it from the strong DMM. Not only does the strong DMM require nonintersection of all ISRFs, but also the same ordering of the k ISRFs for each value of item score x ($x = 1, \ldots, m$). (Equation 29 shows this assumption only for two arbitrary items i and j.)

Finally, Scheiblechner (1995) proposed the isotonic ordinal probabilistic (ISOP) model, which for polytomous items is described by Equation 29 but not by Equation 28. Thus, the ISOP model assumes the same invariant ordering of the k ISRFs across values of x, but these m bundles of k ISRFs each are allowed to intersect with one another (that is, Equation 28 does not hold). The strong DMM, therefore, is a special case of the weak DMM and of the ISOP model, but the weak DMM and the ISOP model are mutually exclusive models.

Alternative models proposed by, for example, Hemker et al. (1996, 1997) and Samejima (1972, 1995) will not be considered because of their limited familiarity compared with most models mentioned here. Bock's (1972) nominal response model is not considered because it was defined for nominal response data whereas all other models mentioned here assume an ordering of the response categories per item.

TABLE 2
Characteristics of ISRFs of 10 IRT Models for Polytomous Items

	Lowest Prob.	Highest Prob.	Slope Across Items	Slope Within Items	Intersection (Across Items)
RSM	0	1	const	const	No
PCM	0	1	const	const	No
G-PCM	0	1	var	const	Yes
OPLM-po	0	1	var	const	Yes
RS-GRM	0	1	var	const	Yes
GRM	0	1	var	const	Yes
MHM-po	≥ 0	≤ 1	var	var	Yes
WEAK DMM	≥ 0	≤ 1	var	var	No
STRONG DMM	≥ 0	≤ 1	var	var	No
ISOP	≥ 0	≤ 1	var	var	Yes

List of Abbreviations:

RSM:	Rating Scale Model
PCM:	Partial Credit Model
G-PCM:	Generalized Partial Credit Model
OPLM-po:	One-Parameter Logistic Model (polytomous items) with Imputed Slopes
RS-GRM:	Rating Scale version of Graded Response Model
GRM:	Graded Response Model
MHM-po:	Monotone Homogeneity Model (polytomous items)
WEAK DMM:	Weak Double Monotonicity Model
STRONG DMM:	Strong Double Monotonicity Model
ISOP:	Isotonic Ordinal Probabilistic Model

Variation of M in Polytomous IRT Models

All 10 IRT models for polytomous item scores discussed have UD, LI, and M in common, and differ in the restrictions imposed on M. Table 2 provides an overview of the models and of characteristics of the ISRFs given M. Compared with Table 1 for IRT models for dichotomous items, the column pertaining to inflection points has been deleted in favor of the distinction between whether or not the slopes of the ISRFs vary across or within items. For the models considered here the following can be seen:

(1) Parametric models have ISRFs with minimum values of 0, whereas nonparametric models allow higher minimum values.
(2) Parametric models have ISRFs with maximum values of 1, whereas nonparametric models allow lower maximum values.
(3) The ISRFs of the rating scale model and the partial credit model have constant slopes across items. All other models in Table 2 allow varying slopes across items. Note however, that for the weak DMM and the strong DMM variation is restricted by the nonintersection of the ISRFs and that

for the ISOP model this restriction pertains to bundles of *ordered* nonintersecting ISRFs across items.

(4) Within items, all parametric models have ISRFs with equal slopes; whereas all nonparametric models allow varying slopes. Note that the structural nonintersection of ISRFs within items restricts the variation in slope.

(5) The rating scale model and the partial credit model (parametric models) and the weak DMM and the strong DMM (nonparametric models) have ISRFs which do not intersect across the items. Note that the ISOP model only allows intersection of ISRFs from different items if they pertain to different item scores.

A Taxonomy of IRT Models

The taxonomy of IRT models for dichotomous items and IRT models for polytomous items shows which models imply SOL (Equation 6), which models imply IIO (Equation 8), and which models imply both SOL and IIO or neither of these properties. The proofs that particular models do or do not have SOL or IIO were given mostly elsewhere (Grayson, 1988; Hemker et al., 1996, 1997; Huynh, 1994; Sijtsma & Hemker, 1998; Sijtsma & Junker, 1996); in the Appendix, we show that the strong DMM (Sijtsma & Hemker, 1998) and the ISOP model (Scheiblechner, 1995) do not imply SOL. This is a new result not proven elsewhere.

Stochastic Ordering

Dichotomous Items. Grayson (1988) proved the following important result for all IRT models for dichotomous items that are UN, LI, and M. As before, assume that $s_1 < s_2$; then

$$g(s_1,s_2;\theta) = \frac{P(X_+ = s_1|\theta)}{P(X_+ = s_2|\theta)} \tag{30}$$

is nondecreasing in θ. Equation 30 expresses that X_+ has *monotone likelihood ratio* (MLR) in θ. All dichotomous IRT models in Table 3 have UN, LI, and M in common, and each of them has the MLR property. The reason MLR is so important is that it *implies* SOL (Equation 6); conversely, SOL does *not* imply MLR (Lehmann, 1959, 1994, p. 74). Therefore, by implication SOL holds for all dichotomous IRT models.

Polytomous Items. Hemker et al. (1996) investigated MLR for polytomous IRT models, and Hemker et al. (1997) investigated SOL for polytomous IRT models (also see the Appendix). MLR implies SOL and, moreover, SOL has more practical relevance. Hemker et al. (1997) did not identify general conditions under which SOL held for polytomous IRT models that are UD and LI. Instead, the results for SOL obtained by Hemker et al. (1997) were with respect to separate models rather than a general class. Table 3 shows that the rating scale

TABLE 3
Presence (+) or Absence (−) of the Properties of Stochastic Ordering of the Latent Trait (SOL) and Invariant Item Ordering (IIO) in IRT Models

Dichotomous Data	SOL	IIO	Polytomous Data	SOL	IIO
1-PLM	+	+	RSM	+	+
2-PLM	+	−	PCM	+	−
3-PLM	+	−	G-PCM	−	−
4-PLM	+	−	OPLM-po	−	−
OPLM	+	−	RS-GRM	−	−
1-PNOM	+	+	GRM	−	−
2-PNOM	+	−	MHM-po	−	−
3-PNOM	+	−	WEAK DMM	−	−
MHM-di	+	−	STRONG DMM	−	+
DMM-di	+	+	ISOP	−	+

List of Abbreviations:

1-PLM:	1-Parameter Logistic Model (Rasch model)
2-PLM:	2-Parameter Logistic Model (Birnbaum model)
3-PLM:	3-Parameter Logistic Model
4-PLM:	4-Parameter Logistic Model
OPLM:	One-Parameter Logistic Model with Imputed Slopes
1-PNOM:	1-Parameter Normal Ogive Model
2-PNOM:	2-Parameter Normal Ogive Model
3-PNOM:	3-Parameter Normal Ogive Model
MHM-di:	Model of Monotone Homogeneity for Dichotomous Data
DMM-di:	Model of Double Monotonicity for Dichotomous Data
RSM:	Rating Scale Model
PCM:	Partial Credit Model
G-PCM:	Generalized Partial Credit Model
OPLM-po:	One-Parameter Logistic Model for Polytomous Items
RS-GRM:	Rating Scale version of the Graded Response Model
GRM:	Graded Response Model
MHM-po:	Model of Monotone Homogeneity (polytomous items)
WEAK DMM:	Weak Double Monotonicity Model
STRONG DMM:	Strong Double Monotonicity Model
ISOP:	Isotonic Ordinal Probabilistic Model

model (Andrich, 1978) and the partial credit model (Masters, 1982) have SOL, but none of the other models has SOL.

Invariant Item Ordering

Dichotomous Items. If two IRFs intersect, the ordering of the probabilities $P_i(\theta)$ is opposite left and right of the intersection point. Because the only requirement for an IIO in a dichotomous IRT model is that the IRFs do *not* intersect, it is easily established that the 1PLM, the 1-parameter normal ogive

model, and the DMM imply an IIO. All other models do not imply an IIO (Table 3) because their IRFs may intersect.

Polytomous Items. For polytomous IRT models, the properties of the CCCs or of the ISRFs together determine whether a particular IRT model implies IIO (Sijtsma & Hemker, 1998). Of the parametric models listed in Table 3, the rating scale model (Andrich, 1978) implies IIO, but none of the other models implies IIO. Also, the strong DMM (Sijtsma & Hemker, 1998) and the ISOP model (Scheiblechner, 1995) imply IIO.

Discussion

Measurement practitioners may prefer to use the total score X_+, rather than the estimate of θ due to the former's familiarity and simplicity, which make it easier to communicate test results to non-specialists or laymen. If the purpose of testing is the ordering of respondents on θ, this is good practice if dichotomous items were used and if the data comply with any of the 10 dichotomous IRT models in Table 3. SOL is a property which holds for all of these models. This is not a self-evident result, especially if one realizes that only in the 1PLM or Rasch model is X_+ a sufficient statistic for estimating θ. However, here a point estimate of θ is obtained, whereas our stochastic ordering result concerns the ordering on θ.

For polytomous IRT models, SOL only holds for the partial credit model and for the rating scale model, which is a special case of the former model, but not for any of the other models. Thus, using X_+ for ordering respondents on θ may not represent the true ordering under most polytomous IRT models. In future research, the degree to which the SOL property is violated under the application envisaged needs to be investigated.

To anticipate such research, and to have some first impressions, we did some preliminary calculations for a standard normal θ and item parameters that seemed fairly representative of testing in practical applications. In the first example, we used the graded response model (Equation 25) with $k = 4$, $m + 1 = 3$; $\alpha_i = 1$ $(i = 1, \ldots, 4)$; $\lambda_{11} = -1$, $\lambda_{12} = 1$; $\lambda_{21} = -\frac{1}{2}$, $\lambda_{22} = \frac{1}{2}$; $\lambda_{31} = -1$, $\lambda_{32} = \frac{1}{2}$; and $\lambda_{41} = -\frac{1}{2}$, $\lambda_{42} = 1$. The second example was different from the first example in that $\alpha_1 = \alpha_2 = \alpha_4 = 1$ and $\alpha_3 = 2$. For both examples, we calculated with great accuracy the probabilities $P(\theta > t \mid X_+ = s)$ with $t = -3, -2, -1, 0, 1, 2, 3$; and $s = 0, \ldots, 8$. For each t, $P(\theta > t \mid X_+ = s)$ increased in s (Equation 6); thus, SOL was valid here. Similar examples led to the same conclusion. Given these positive results, it seems worthwhile in future research to study the conditions under which SOL holds.

Two final remarks about the use of X_+. First, using X_+ rather than θ means that the metric scale properties of θ are not used in practice when communicating results to laymen. This is not problematic since, in ordinary practice, orderings of respondents rather than distances between respondents are important and, moreover, the interpretation of distances in terms of psychological

quantities/traits would be quite troublesome. Besides, since θ is used in all basic research, the metric properties of the θ scale can be exploited there for equating scales, building item banks, and adaptively testing respondents. Second, the reliability of an ordering may be expressed by Kendall's rank correlation explicitly using information from concordant, discordant and tied respondent pairs instead of by the product-moment correlation which also uses distance information. Sijtsma and Molenaar (1987) noted that the conclusions from rank and product-moment correlations are almost equivalent. Moreover, basic research will preferably use θ and, if available, the Fisher information function, which provides information on the accuracy of the maximum likelihood estimate of θ, conditional on θ.

If a model does not imply an IIO, results pertaining to item ordering are more difficult to interpret and for many applications the functioning of a test may not be understood completely. For example, if, contrary to expectation, the items were to have a different ordering for boys and girls, this result would call for additional research aimed at explaining the different item orderings. Such research could, for example, involve the use of DIF methods (e.g., Holland & Wainer, 1993). Of course, models not implying an IIO can be used to construct tests, but if an IIO is considered important for the application envisaged, this property has to be investigated separately in addition to the fit investigation of the IRT model to the data.

It may be noted that if an IRT model for dichotomous items implies intersecting IRFs resulting for k items in, say, K intersection points in total, the θ scale is divided into $K + 1$ exhaustive and mutually exclusive intervals. A particular item ordering according to $P_i(\theta)$ $(i = 1, \ldots, k)$ exists for each of these intervals. For example, a pair of IRFs $(\alpha_i \neq \alpha_j)$ from the 2PLM has one intersection point; thus, k IRFs with k different αs have $\frac{1}{2}k(k-1)$ intersection points, and $\frac{1}{2}k(k-1)+1$ intervals are defined, each characterized by a unique item ordering according to $P_i(\theta)$ $(i = 1, \ldots, k)$. For polytomous IRT models not implying an IIO a similar line of reasoning can be given.

Thus, if a model does not imply an IIO, we know that several different item orderings exist and, moreover, that orderings may be much different from one another. Sijtsma and Junker (1996) discussed nonparametric methods for investigating whether IIO holds for a test based on dichotomous items, and Sijtsma and Hemker (1998) discussed a nonparametric method for investigating IIO in case of polytomous items.

Appendix

We give two numerical examples which show that the strong DMM does not imply stochastic ordering of θ by X_+ (thus far abbreviated SOL). Because the strong DMM is a special case of the ISOP model, by implication these examples also demonstrate that the ISOP model does not imply SOL. The examples are elaborations of an example given by Hemker et al. (1997), which showed that the MHM does not imply stochastic ordering of θ by item score X_i. Since we

distinguish stochastic ordering of θ by X_+ and stochastic ordering of θ by X_i, for clarity we will use the abbreviations SO by X_+ and SO by X_i.

Example 1. Let $0 \leq \theta \leq 1$; and consider two items, i and j, with three ordered answer categories (X_i, $X_j = 0, 1, 2$). The ISRFs are defined as

$$P(X_i \geq 1 \mid \theta) = \begin{cases} 3\theta, & \text{if } 0 \leq \theta < \frac{1}{4}; \\ \frac{2}{3} + \frac{1}{3}\theta, & \text{if } \frac{1}{4} \leq \theta \leq 1; \end{cases}$$

$$P(X_i \geq 2 \mid \theta) = \begin{cases} 2\theta, & \text{if } 0 \leq \theta < \frac{1}{4}; \\ \frac{1}{4} + \theta, & \text{if } \frac{1}{4} \leq \theta < \frac{1}{2}; \\ \frac{1}{2} + \frac{1}{2}\theta, & \text{if } \frac{1}{2} \leq \theta \leq 1; \end{cases}$$

$$P(X_j \geq 1 \mid \theta) = \begin{cases} \frac{4}{3}\theta, & \text{if } 0 \leq \theta < \frac{1}{4}; \\ \frac{1}{9} + \frac{8}{9}\theta, & \text{if } \frac{1}{4} \leq \theta \leq 1; \end{cases}$$

and

$$P(X_j \geq 2 \mid \theta) = \theta, \text{ if } 0 \leq \theta \leq 1.$$

These ISRFs are nondecreasing and, moreover, it can be checked that

$$P(X_i \geq 1 \mid \theta) \geq P(X_i \geq 2 \mid \theta) \geq P(X_j \geq 1 \mid \theta) \geq P(X_j \geq 2 \mid \theta). \qquad \text{(A1)}$$

Thus, the ISRFs do not intersect, and they comply with Equation 29. In combination with nondecreasingness, these results imply the strong DMM. The combination of Equation A1 and Equation 13 (IRF expressed as sum of ISRFs) implies Equation 8 (IIO; Sijtsma & Hemker, 1998).

Consider a three-point distribution of θ, with $P(\theta = \frac{1}{4}) = P(\theta = \frac{1}{2}) = \frac{1}{4}$ and $P(\theta = 1) = \frac{1}{2}$. Using this distribution, it can be shown that SO by X_i does *not* hold (Hemker et al., 1997); it also can be shown, however, that SO by X_j *does* hold. To investigate whether SO by X_+ (defined as $X_+ = X_i + X_j$) holds, we calculate, for $x_+ = 0, 1, 2, 3, 4$, the probabilities $P(\theta > \frac{1}{4} | X_+ = x_+)$, which yields 0.31, 0.20, 0.50, 0.44 and 0.95, respectively. Thus $P(\theta > \frac{1}{4} | X_+ = x_+)$ is not nondecreasing in x_+, meaning that SO by X_+ does not hold.

Example 2. If item i has SO by X_i *and* item j has SO by X_j (in Example 1 only item j had this SO property), and if the strong DMM holds, then SO by X_+ ($X_+ = X_i + X_j$) need not be implied. The same definition of item j and the same

three-point distribution of θ as in Example 1 are used. Consider a new item i with three ordered answer categories ($X_i = 0, 1, 2$) and with ISRFs

$$P(X_i \geq 1 \mid \theta) = \begin{cases} \frac{2}{3}\theta, & \text{if } 0 \leq \theta < \frac{1}{2}; \\ \frac{4}{3}\theta - \frac{1}{3}, & \text{if } \frac{1}{2} \leq \theta \leq 1; \end{cases}$$

and

$$P(X_i \geq 2 \mid \theta) = \begin{cases} \frac{33}{50}\theta, & \text{if } 0 \leq \theta < \frac{1}{2}; \\ \frac{67}{50}\theta - \frac{17}{50}, & \text{if } \frac{1}{2} \leq \theta \leq 1. \end{cases}$$

It can be checked that these ISRFs are nondecreasing and, moreover, that

$$P(X_j \geq 1 \mid \theta) \geq P(X_j \geq 2 \mid \theta) \geq P(X_i \geq 1 \mid \theta) \geq P(X_i \geq 2 \mid \theta). \qquad \text{(A2)}$$

It can be shown that, given the choice of the θ distribution, SO by X_i holds. For $x_+ = 0, 1, 2, 3, 4$, we have that $P(\theta > 1/4 | X_i + X_j = x_+) = 0.35, 0.35, 0.60, 0.59$, and 0.98, respectively. Thus $P(\theta > 1/4 | X_i + X_j = x_+)$ is not nondecreasing in x_+, meaning that SO by X_+ does not hold.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Akkermans, L.M.W. (1998). *Studies on statistical models for polytomously scored items*. PhD Thesis, University of Twente, The Netherlands.

Akkermans, L.M.W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62*, 569–579.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*, 101–119.

Bartholomew, D. J. (1996). *The statistical approach to social measurement*. San Diego, CA: Academic Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores* (pp. 396–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391–404.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*, 679–693.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331–347.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59*, 77–79.

Irtel, H. (1995). An extension of the concept of specific objectivity. *Psychometrika, 60*, 115–118.

Knott, M., & Albanese, M. T. (1993). Conditional distributions of a latent variable and scoring for binary data. *Revista, Brasiliva de Probabilidade e Estatistica, 6*, 171–188.

Lehmann, E. L. (1959, 1994). *Testing statistical hypotheses*. New York: Wiley/Chapman & Hall.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph No. 7*, Psychometric Society.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Meijer, R. R., Sijtsma, K., & Smid, N. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*, 283–298.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 292–299.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417–430.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden, 3*(8), 145–164.

Molenaar, I. W. (1983). *Item steps. (Heymans Bulletin 83-630-EX)*. Groningen, The Netherlands: University of Groningen.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 3–14). New York: Springer.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351–363.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monography, No. 18.*

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general guarded response method. *Psychometrika, 60*, 549–572.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60*, 281–304.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3–31.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63*, 183–200.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79–105.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*, 79–97.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191–199.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory.* PhD Thesis, University of Amsterdam, The Netherlands.

Verhelst, N. D., & Glas, C.A.W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–237). New York: Springer-Verlag.

Verhelst, N. D., & Verstralen, H.H.F.M. (1991). *The partial credit model with non-sequential solution strategies.* Arhhem, The Netherlands: Cito National Institute for Educational Measurement.

Authors

KLAAS SIJTSMA is a professor of Psychological Research Methods in the Department of Research Methodology, FSW, Tilburg University, PO Box 90153, 5000 Tilburg, The Netherlands; k.sijtsma@kub.nl. He specializes in research methodology, applied statistics, and psychometrics.

BAS T. HEMKER is a senior researcher at the CITO National Institute for Educational Measurement, PO Box 1034, 6801 MG Arnhem, The Netherlands; bas.hemker @cito.nl. He specializes in educational measurement and psychometrics.

Received July 1998
Revision Received June 1999
Accepted January 2000